Centromere-proximal differentiation and speciation in *Anopheles gambiae*

Aram D. Stump*, Meagan C. Fitzpatrick*, Neil F. Lobo*, Sékou Traoré[†], N′Fale Sagnon[‡], Carlo Costantini[‡§], Frank H. Collins*, and Nora J. Besansky*[¶]

*Center for Tropical Disease Research and Training, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; [†]Département d'Epidémiologie des Affectations Parasitaires, Faculté de Médecine, de Pharmacie, et d'Odonto-Stomatologie, Université du Mali, Boite Postale 1805, Bamako, Mali; [‡]Centre National de Recherche et Formation sur le Paludisme, Ouagadougou, Burkina Faso; and [§]Unité de Recherche 016, Caractérisation et Contrôle des Populations de Vecteurs, Institut de Recherche pour le Développement, Ouagadougou, Burkina Faso

Communicated by Margaret G. Kidwell, University of Arizona, Tucson, AZ, September 19, 2005 (received for review July 11, 2005)

The M and S molecular forms of *Anopheles gambiae* are undergoing speciation as they adapt to heterogeneities in the environment, spreading malaria in the process. We hypothesized that their divergence despite gene flow is facilitated by reduced recombination at the centromeric (proximal) end of the X chromosome. We sequenced introns from 22 X chromosome genes in M and S from two locations of West Africa where the forms are sympatric. Generally, in both forms nucleotide diversity was high distally, lower proximally, and very low nearest the centromere. Conversely, differentiation between the forms was virtually zero distally and very high proximally. Pairwise comparisons to a close relative, the sibling species *Anopheles arabiensis*, demonstrated uniformly high divergence regardless of position along the X chromosome, suggesting that this pattern is not purely mechanical. Instead, the pattern observed for M and S suggests the action of divergent natural selection countering gene flow only at the proximal end of the X chromosome, where recombination is reduced. Comparison of sites with fixed differences between M and S to the corresponding sites in *A. arabiensis* revealed that derived substitutions had been fixed in both forms, further supporting the hypothesis that both have been under selection. These derived substitutions are fixed in the two West African samples and in samples of S from western and coastal Kenya, suggesting that selection occurred before the forms expanded to their current ranges. Our findings are consistent with a role for suppressed genetic recombination in speciation of *A. gambiae*.

gene flow | natural selection | polymorphism | recombination | reproductive isolation

Reduced recombination contributes to the persistence of species or emerging species in the face of gene flow (1, 2). Despite some interbreeding and hybrid formation, species-specific regions of the genome can be preserved from introgression and homogenization of different genetic backgrounds if they are not subject to crossing over. Recent elaborations of this concept in organisms as diverse as humans, sunflowers, fruit flies, and mosquitoes invoke chromosomal inversions because recombination is effectively suppressed between the break points of chromosomal inversion heterozygotes (3–6). If captured by inversions, genes involved in assortative mating and species-specific ecological adaptations would remain associated longer relative to the case of free recombination. For pairs of species that have diverged in parapatry or sympatry, these models predict (*i*) that they are likely to differ by fixed inversions, (*ii*) that genes involved in reproductive isolation and species-specific adaptations should preferentially map to these inversions, and (*iii*) that significantly greater genetic divergence will accumulate in rearranged versus colinear regions. These predictions seem to be upheld for the group of mosquito sibling species known as the *Anopheles gambiae* complex (7, 8) but not for *A. gambiae* itself.

A. gambiae, the primary vector of malaria in subSaharan Africa, is undergoing speciation. The incipient species (designated molecular forms M and S) are defined by fixed sequence differences in the X-linked ribosomal DNA locus (ref. 9 and references therein), but these species do not differ by any known fixed inversion difference. The S form occurs across subSaharan Africa, where it exploits typical rain-dependent *A. gambiae* breeding sites. The M form is found primarily west of the 20°E parallel and seems to be associated with more permanent larval breeding sites, including rice fields and reservoirs, that extend the reach of this vector (and, thus, malaria) into more arid regions and seasons. Nevertheless, where sympatric throughout most of west Africa (figure 1 of ref. 9), the two forms can be sampled as adults from the same houses, where females of both forms feed preferentially on human blood. Wild females are estimated to mate with males of the other form at a rate of only ≈1% (10), indicating strong assortative mating. In the laboratory, hybrids of both sexes are viable and fertile, but naturally occurring M/S hybrids are rare (e.g., 6 of 8,000, ≪1%) (9).

Consistent with small amounts of ongoing gene flow, most previous investigations of genetic diversity in *A. gambiae* have reported no or only slight differentiation between the two forms (11, 12). Aside from sequences adjacent to the centromere of chromosome 2L and a small region of 2R (13, 14), the most significant differentiation found to date has been at the centromeric end of the X chromosome (14, 15). Nine microsatellite loci in an ≈5-megabase region of subdivisions 5B-6 have F_{ST} values higher than any found elsewhere on the X or on other chromosomes (mean $F_{ST} = 0.220$, with values at a locus as high as 0.468) (15). In this same region, there is a nearly fixed difference in the insertion of a *Maque* transposable element at one locus, and strong frequency differences at others (16). The proximal end of the X chromosome is also the location of the ribosomal DNA locus, which defines the molecular forms. This pattern (strong differentiation across an ≈5-megabase region of the genome and none in other regions) suggests that the divergence of the molecular forms has been driven by selection on a limited number of loci, consistent with a divergence-with-gene-flow model (17). Our working hypothesis is that at some time in the past the ancestral population started to specialize on different larval habitats, either in parapatry or in sympatry, with positively selected ecological adaptations mapping to loci at the centromeric end of the X in one or both of the populations. Through time, some degree of reproductive isolation has arisen in association with these genetic backgrounds, conferred by loci that map to the same region. Because isolation is incomplete, hybrids afford some gene flow between the two populations, which therefore maintain homogeneity over much of their genomes. However, the strength of ongoing selection to maintain alternate ecological specializations overcomes

Conflict of interest statement: No conflicts declared.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ101280–DQ102293 and DQ102295–DQ102333).

[¶]To whom correspondence should be addressed at: Department of Biological Sciences, 317 Galvin Life Sciences Building, University of Notre Dame, Notre Dame, IN 46556-0369. E-mail: nbesansk@nd.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0508161102

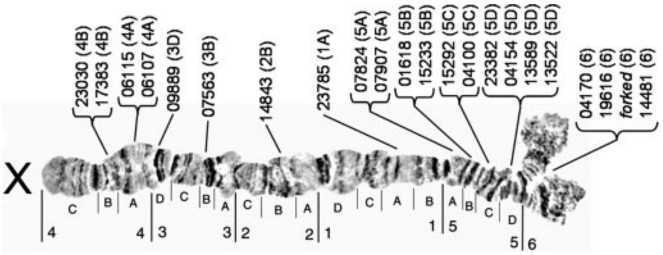

Fig. 1. Relative location of loci on a cytogenetic map of the *A. gambiae* X chromosome, oriented with the centromere on the right. Numbering of divisions and subdivisions is relative to an arbitrary chromosomally standard reference (33).

gene flow at the loci conferring those traits. We further propose that natural selection is aided in this process by a chromosomal feature equivalent in its effect to fixed inversion differences: reduced levels of recombination typical of pericentromeric regions (18).

The working hypothesis predicts that positive selection and associated genetic hitchhiking should reduce diversity and increase divergence and levels of linkage disequilibrium in candidate regions at the proximal end of the X chromosome but not outside this region. Our previous survey of microsatellite loci on the X chromosome found high divergence but little reduction in polymorphism or linkage disequilibrium at most loci located in the centromere-proximal divisions 5–6 (15), likely because of the high mutation rate of microsatellites, which allowed polymorphism to recover since the selective sweeps. Accordingly, more slowly mutating intron sequences may still bear the expected signature of positive selection. To test these predictions, we determined intron sequence variation at 22 genes distributed along the length of the X chromosome in M and S populations of *A. gambiae*. To ensure that the patterns of population differentiation would not be specific to a particular locality, specimens came from two locations in West Africa where M and S are sympatric. S form specimens from two locations in East Africa were added to determine whether alleles characteristic of the S form in West Africa are also found in areas where the M form is absent. Sequences from the sibling species *Anopheles arabiensis* were included to ensure that patterns of polymorphism and differentiation in *A. gambiae* were not reflected in other species due solely to mechanical or mutational constraints at the proximal end of the X chromosome.

Materials and Methods

A. gambiae adults were sampled from Goundri, Burkina Faso, in September 2001 by pyrethrum spray catch. In Mali, collections were by manual aspiration of adults from Banambani in July 1997 and from Bancoumana and Moribabougou in June 2004. In Kenya, collections were made by using pyrethrum spray in Asembo (western Kenya) in November 2001 and by manual aspiration in Jego (coastal Kenya) in May 1987. Specimens were identified morphologically as belonging to the *A. gambiae* sibling species complex (19), and only males were chosen to facilitate direct sequencing of the hemizygous X chromosome. DNA was isolated by using DNeasy Tissue kits (Qiagen, Valencia, CA). Mosquitoes were identified to species and molecular form with a ribosomal DNA-based PCR-restriction fragment length polymorphism assay (20).

A panel of introns was chosen to span the length of the X chromosome, with emphasis on divisions 5 and 6 (Fig. 1). Individual introns were identified by using gene predictions available in the Ensembl *A. gambiae* genome browser (www.ensembl.org/Anopheles_gambiae). One intron in our panel was from a gene that escaped annotation but had been identified as a homolog to the *Drosophila melanogaster* gene *forked* in the course of a previous study (16); intron/exon structure for this gene was predicted by comparative sequence alignment. PRIMER3 (21) was used to design primers in flanking exons to amplify $\approx$500-bp introns, where possible. Alternatively, one of the primers was designed inside longer introns, and, in one case, primers flanked three smaller introns (Table 3, which is published as supporting information on the PNAS web site). We adopted the following naming convention: the last five digits of the Ensemble gene ID followed by the cytogenetic map location in parentheses.

PCR was performed in a GeneAmp 9600 thermal cycler (Applied Biosystems). Each 50-μl reaction contained 25 pmol of each primer, each dNTP at 0.2 mM, 1.5 mM $MgCl_2$, 5 units of Taq polymerase, and 1 μl of a 1:4 dilution of template DNA extracted from a single mosquito. Cycling conditions were 94°C denaturation for 5 min followed by 40 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 1 min, with a final 72°C extension of 5 min. PCR products were separated in a 1% agarose gel, excised, and isolated from the gel matrix by using gel extraction kits (Qiagen). Sequencing was carried out by using the Applied Biosystems PRISM Dye Terminator Cycle Sequencing kit and an Applied Biosystems 3700 sequencer, with internal primers as necessary to achieve complete coverage of both strands (Table 3). Sequences were assembled and verified by using SeqMan II (DNASTAR, Madison, WI). The sequences have been deposited in the GenBank database under accession numbers DQ101280–DQ102293 and DQ102295–DQ102333.

Sequence alignments performed using CLUSTALX (22) were inspected by eye, and minor modifications were made as necessary to maximize similarity. To prevent polymorphism statistics from being skewed by the inclusion of different lengths of nearly monomorphic exon sequences, only intron sequences were analyzed. Sequences of M and S obtained from two locations in West Africa were included in analyses of polymorphism and linkage disequilibrium, except for one aberrant M sequence at locus 07824(5A) (see *Results*). The S form sequences from East Africa were not included to avoid biasing the results by pooling heterogeneous samples in one form but not the other. Basic sequence statistics including nucleotide diversity (π) and expected heterozygosity (θ), pairwise tests of linkage disequilibrium between parsimony informative sites, minimum number of recombination events, Tajima's D, and measures of sequence divergence (D_a and F_{ST}) were calculated with DNASP 4.00 (23). F_{ST} values were tested for significance by using 10,000 permutations in ARLEQUIN 2.0 (24). DNASP was also used to run 10,000 coalescent simulations to evaluate the probability of the observed nucleotide diversity in chromosomal divisions 5 and 6 conditioned on expected heterozygosity across the total data set.

A maximum-likelihood multilocus test of the standard neutral model based on the Hudson–Kreitman–Aguade test was applied to polymorphism data for 22 genes from M or S by using divergence data from *A. arabiensis* (25). With this method, we explicitly tested for selection in chromosomal divisions 5 and 6, assuming loci in divisions 1–4 to be evolving neutrally. Markov chain lengths of at least 10^6 were required for convergence, as assessed by eight independent runs per model, with different random number seeds.

Neighbor-joining trees were constructed and tested with 10,000 bootstrap replicates by using MEGA 2.1 (26). For gene tree construction, sequences of S form from East Africa, *A. arabiensis* (where possible), and the corresponding reference sequence from the *A. gambiae* PEST genome were included along with West African M and S sequences.

Results

Intron sequences (each $\approx$500 bp) from 22 predicted genes on the X chromosome were determined for *A. gambiae* M and S (Table 3). The relative distribution of these loci along the X chromosome is shown in Fig. 1. For all 22 loci, sequences were determined from at least 15 M and 10 S from West Africa and at least two S from East Africa. In addition, sequences were obtained from at least two specimens from the sibling species, *A. arabiensis*, for 16 of 22 loci

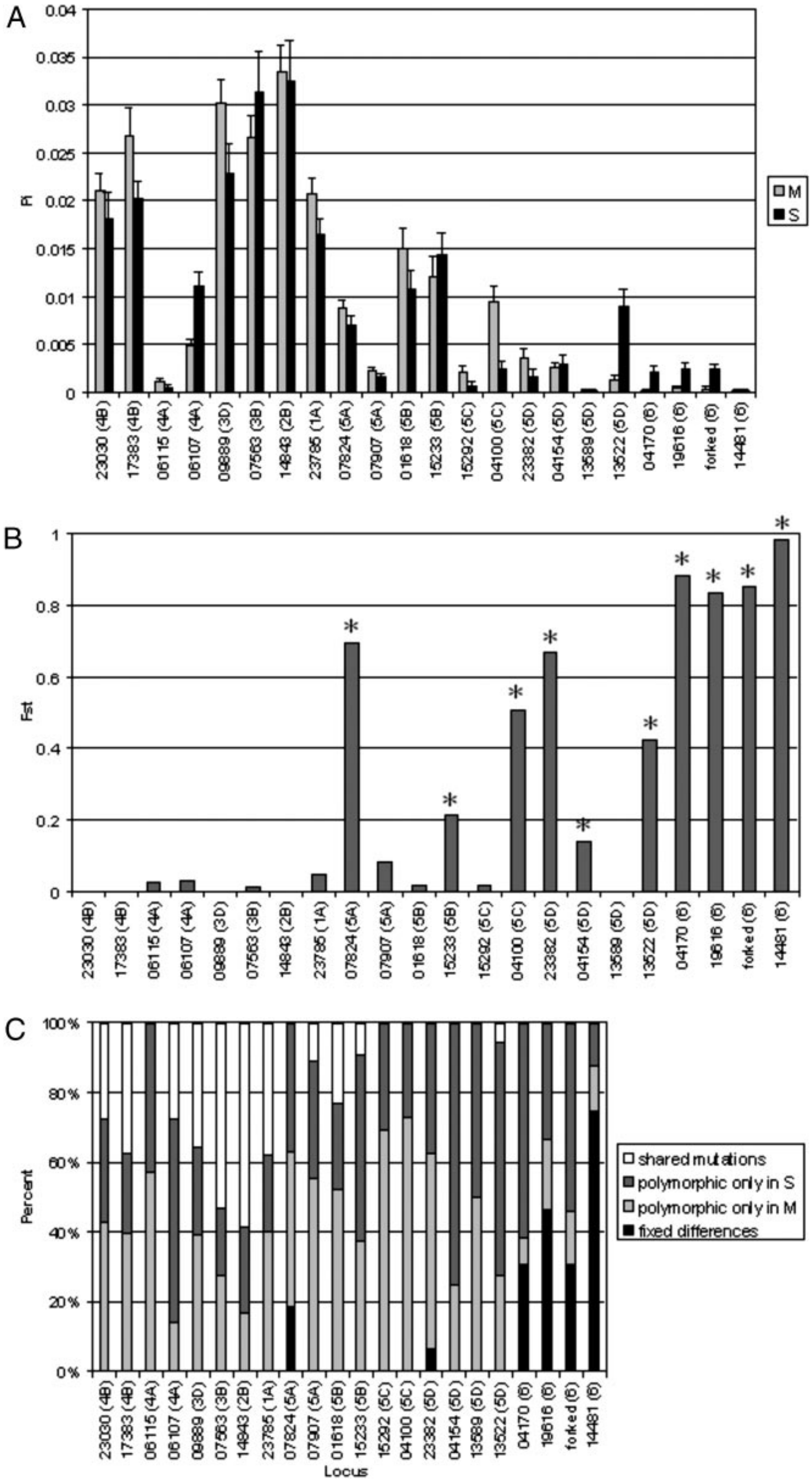

Fig. 2. Polymorphism and divergence in West African samples of *A. gambiae* M and S. The *x* axis showing the loci that were surveyed is not drawn to scale. (*A*) Nucleotide diversity. Error bars indicate one standard deviation. (*B*) Differentiation. Asterisks indicate significant F_{ST} values after Bonferroni correction. (*C*) Proportion of fixed differences and shared or exclusive polymorphisms. Polymorphisms include insertion/deletions.

(Table 4, which is published as supporting information on the PNAS web site).

Polymorphism. Mean nucleotide diversity in M and S was high across the eight loci in cytogenetic map divisions 1–4 (π = 0.0208 and 0.0191 for M and S, respectively), and lower by 4-fold across the 14 loci in divisions 5 and 6 (π = 0.0043) (Fig. 2*A* and Table 5, which is published as supporting information on the PNAS web site), most notably in division 5D-6 closest to the centromere. The probability of observing such reduced diversity at a locus is low, as shown by coalescent simulations without recombination, conditioned on an expected heterozygosity value representing the average across all 22 loci (given $\theta = 0.0135$ for M and S, $P_{(\pi \leq 0.0043)} = 0.038$). Considering only division 6, the probability of observing π at $\leq$0.0003 in M and π at $\leq$0.0019 in S is even smaller (P = 0.0000 and 0.0014, respectively). For the sibling species *A. arabiensis,* there was no pattern of reduced diversity in divisions 5 and 6 relative to the rest of the chromosome. Instead, diversity was uniformly low for all 16 loci ($\pi = 0.0040$) (Table 5). Although variance on these estimates is high given small sample size, the trend of uniformly low diversity along the entire chromosome is consistent with more robust microsatellite-based values for this taxon, which were assessed from the same samples (15).

Estimates of R_m, the minimum number of recombination events at a locus (27) within each form, indicated fewer events toward the proximal end of the X, and no recombination was detected in division 5D-6 (Table 5). This and other summary statistics are affected by low nucleotide diversity; more direct evidence from laboratory genetic crosses confirms suppressed recombination in divisions 5 and 6 (M. Pombi, A.D.S., N.J.B. and A. della Torre, unpublished data). Intragenic tests of linkage disequilibrium within forms revealed some significant associations between pairs of sites from loci outside of divisions 5 and 6 (data not shown). However, most of such tests were not possible at the proximal end of the X chromosome because of low numbers of polymorphic sites, most of which were singletons (Table 5). Tests of intergenic linkage disequilibrium were conducted within each form by using concatenated data sets spanning divisions 5 and 6. Although several pairs of sites were significantly associated between genes by Fisher's exact test (30 in M and eight in S overall), no comparison remained significant after Bonferroni correction.

Divergence. Between M and S, the number of net nucleotide substitutions per site, D_a (28), and F_{ST} values based on nucleotide divergence (29) varied according to location on the X chromosome (Fig. 2*B*). Across eight loci in divisions 1–4, D_a was zero and F_{ST} was 0.018. The corresponding values increased to 0.004 and 0.273 for 10 loci in division 5 and further increased to 0.010 and 0.874 for four loci in division 6. In contrast, for all 16 loci available for comparison between *A. arabiensis* and M or S, F_{ST} values ranged from 0.76–0.99 along the length of the X (data not shown). Although the magnitude of these values may be biased by small sample size of *A. arabiensis*, the trend is consistent with significant interspecific divergence across the X chromosome based on microsatellite variation measured from 50 (each) of the same samples (15).

Following the same trend of increasing differentiation moving proximally along the X chromosome, polymorphisms were shared between M and S at seven of eight loci surveyed in divisions 1–4, at four of the 10 loci in division 5, and none in division 6 (Fig. 2*C*). Conversely, fixed differences between the forms were absent in divisions 1–4, present in two of 10 genes in division 5, and in all loci surveyed in division 6 (Fig. 2*C*).[‖] At each of the six loci bearing fixed (or nearly fixed) differences between M and S, the character state typical of M matched the reference *A. gambiae* PEST genome, and the character state typical of S from West Africa was shared by all East African samples of S, including those from coastal Kenya, east of the Rift Valley. Accordingly, neighbor-joining trees reconstructed from any locus in division 6 showed reciprocal monophyly of M and S but no partitioning of variation within forms by geographic origin. PEST sequences always clustered within M clades. Outside division 6, there was little or no resolution within *A. gambiae*, but *A. gambiae* and *A. arabiensis* sequences were invariably monophyletic (data not shown).

Table 1. Sites with fixed differences between M and S forms of *A. gambiae* compared to *A. arabiensis*

Locus	Ensembl position, MOZ2 assembly	M form	S form	*A. arabiensis*
07824(5A)	15241506	C	T	T
	15241509	C	A	A
	15241634	A	G	G
	15241502–3: +8	Absent	Present	Present
	15241510–11: +10	Absent	Present	Present
04170(6)	20015483	T	A	T
	20015631	G	T	T
	20015945	T	G	G
	20015966	A	C	C
forked(6)	21592987	T	C	T
	21593017	A	C	C
	21593418	A	T	A
	21593436	A	T	A
14481(6)	21853965	A	G	G
	21854051	T	G	T
	21854097	A	T	A
	21854339	T	G	T
	21854348	C	T	T
	21854460	T	C	T

As an outgroup to populations of *A. gambiae*, *A. arabiensis* potentially allows inference of the ancestral state at sites with fixed differences between M and S. Table 1 compares sites with fixed differences between molecular forms to the corresponding *A. arabiensis* site at four loci. For one locus, 07824(5A), the S form carried the presumed ancestral state at each of four sites. However, for the other three loci, 04170(6), *forked*(6), and 14481(6), both forms alternately carried ancestral and derived states at each locus (Table 1).

Tests of Neutrality. Tajima's D statistic is based on the difference between two estimates of variation, θ and π, that should be equal under neutral evolution. Negative values of D indicate an excess of low-frequency polymorphisms, consistent with population expansion or positive selection. Across nearly all loci, values of Tajima's D were negative in both forms, but significantly negative values were obtained only at three loci near the base of the X chromosome: locus 15292(5C) in M and loci 23382(5D) and 04154(5D) in S.

The Hudson–Kreitman–Aguade test (30) is based on a prediction of the neutral model that states that levels of intraspecific polymorphism and interspecific divergence should be correlated. This expectation was evaluated with multilocus data by a maximum likelihood approach (25) that allowed explicit tests of selection at loci in divisions 5 and 6, against reference loci in divisions 1–4 assumed to be neutral. The likelihood-ratio test was used to compare a neutral model (none of the 22 loci under selection) to models in which all (10) or some (six) loci in divisions 5 and 6 are under selection. The subset of six loci in divisions 5 and 6 were chosen based on estimates of the selection parameter k that were consistently <1 among multiple runs, suggesting decreased diversity due to selection. Eight sets of tests were conducted per model with selection, using polymorphism data from each form and divergence estimates from *A. arabiensis*. For both forms, a model that allows selection on a subset of loci in divisions 5 and 6 fits the data significantly better than the neutral model (Table 2).

Time of Hitchhiking. If locus 14481(6) is taken as the most closely linked to a target of selection in M and S (not necessarily the same target), a rough approximation of the time back to the most recent hitchhiking event can be inferred based on the amount of nucleotide variation present on the hitchhiked haplotypes (31). Specifically, the time back to a selective sweep, t, can be estimated by $S/(n\mu)$, where S, n, and μ are the number of segregating sites, the

[‖] There was one exception to these fixed differences. At locus 07824 (5A), one M sequence matched S instead of another M but only at that locus. This finding was confirmed by repeated PCR and sequencing.

Table 2. Maximum-likelihood-ratio Hudson–Kreitman–Aguade test of selection at the base of the X chromosome

Form	Model	ln *L*	Likelihood-ratio statistic (df)	*P* value
M	A. Neutral	−102.3	—	—
	B. Selection on all loci in division 5 and 6	−94.2	A vs. B, 16.2 (10)	NS
	C. Selection on six loci in division 5 and 6	−92.9	A vs. C, 18.8 (6)	<0.01
S	A. Neutral	−102.3	—	—
	B. Selection on all loci in division 5 and 6	−95.6	A vs. B, 13.4 (10)	NS
	C. Selection on six loci in division 5 and 6	−95.0	A vs. C, 14.6 (6)	<0.05

Model C for form M includes loci 07824(5A), 13589(5D), 13522(5D), 04170(6), *forked*(6), and 14481(6). Model C for form S includes loci 07824(5A), 07907(5A), 04100(5C), 13589(5D), *forked*(6), 14481(6). NS, not significant; —, not applicable.

number of sequences, and the mutation rate per sequence per year, respectively. Using 1.1×10^{-8} as the per-nucleotide mutation rate (32) and $S = 1$ for both forms (Table 5), the sweeps in M and S are estimated to have occurred relatively recently, within the last ≈8,000–9,000 years.

Discussion

The diversification of the African malaria vector *A. gambiae* into ecologically distinctive nonpanmictic populations known as M and S may have been associated with adaptation to a changing environment brought about by the development of agriculture within the past 10,000 years (33). If so, signatures of selection may remain in the *A. gambiae* genome and ultimately guide the discovery of genes underlying adaptive divergence and reproductive isolation. We hypothesized that one such signature was located on the X chromosome near the centromere. In a previous test of this hypothesis using microsatellites, we indeed found significant differentiation between M and S at the base of the X chromosome in cytogenetic divisions 5 and 6 (15). However, the inference of selection was weak because in both forms there was no apparent reduction in microsatellite polymorphism in these divisions relative to the rest of the X chromosome. Moreover, our samples were limited to one village in Burkina Faso, West Africa. In the present study, DNA sequencing of identical and thus directly comparable samples from Burkina Faso revealed not only significant differentiation between M and S in divisions 5 and 6 but also greatly reduced polymorphism in this region, consistent with selection on one or more loci located in the pericentromeric region of the X chromosome. Importantly, these results were mirrored in M and S samples from another part of West Africa and in S from East Africa on both sides of the Rift Valley. The conclusion that selection and not demography is responsible for this pattern is supported by the sharp contrast between the centromere-proximal divisions 5 and 6 and the more distal divisions 1–4, in which polymorphism levels were relatively high and divergence was essentially zero. Further support derives from the Hudson–Kreitman–Aguade test of neutrality in divisions 5 and 6, indicating significantly reduced polymorphism relative to divergence. That significantly large negative values from Tajima's *D* were limited to only two loci in S and one in M in divisions 5 and 6 is likely due to a lack of power, owing to the paucity of polymorphism in this region: the footprint of selection. For the same reason, tests of linkage disequilibrium lacked power or could not be performed. Considered together with microsatellite data (15), the evidence implicates positive selection, as discussed next.

The region of the X chromosome showing reduced polymorphism encompasses at least 5 megabases. Both its physical extent and its location near the centromere are consistent with the contribution of reduced recombination, as observed in laboratory genetic crosses of *A. gambiae* M (M. Pombi, A.D.S., N.J.B., and A. della Torre, unpublished data). A positive correlation of recombination and polymorphism can be explained equally well by alternative models of selection: negative (background) or positive (hitchhiking). Under hitchhiking, the episodic sweep of advantageous alleles through the population removes linked neutral variation (34). Under background selection, the continuous removal of deleterious alleles also removes linked neutral variation from the population (35). However, the significant skew in the mutation frequency spectrum seen in this study is unlikely under background selection (36), as is the absence of a similar pattern of reduced diversity restricted to the pericentromeric region of the very closely related sibling species, *A. arabiensis*. An additional argument is based on the theoretical prediction that these processes can be distinguished by their effects on genetic markers that evolve at different rates. The equilibrium dynamics of background selection predict a positive correlation between recombination and genetic variation irrespective of mutation rate. Because the sporadic dynamics of hitchhiking allow a faster recovery of heterozygosity for markers that evolve more rapidly (37–39), no such correlation is expected for microsatellite markers. Mutation rates for SNPs and microsatellites have not been estimated in *A. gambiae*; however, based on evidence from a variety of other organisms, it is likely that microsatellites evolve faster than SNPs by at least three orders of magnitude (39). Therefore, reduced SNP but not microsatellite diversity in the same samples strongly suggests that positive selection rather than background selection is responsible, although the latter cannot be ruled out. Despite the relatively low microsatellite mutation rate of 5.1×10^{-6} in *D. melanogaster* (40), it was estimated that these markers should recover diversity within 1,000 years after a selective sweep (41). Given a comparable number of generations per year in *A. gambiae* (≈10–12), a similar rate of recovery of diversity is plausible. Consistent with this explanation, rough estimates of time since selective sweeps in M and S exceed 1,000 years, which is distant enough to allow recovery of microsatellite diversity at hitchhiking loci and within the time frame of the origins of agriculture.

Although M and S apparently do not differ by fixed chromosomal inversions, they are strongly differentiated in a pericentric region that experiences reduced levels of recombination and may be serving an analogous role in speciation. Extension of the chromosomal speciation hypothesis to the pericentric region of the X chromosome explains several observations. First, our data indicate that there has been divergent positive selection in M and S, not just selection in one form. This finding is a prediction of chromosomal speciation models but not alternative models in which one lineage retains ancestral characteristics after cladogenesis. The phenomenon of divergent natural selection on M and S likely explains the results in Table 1, where, relative to *A. arabiensis*, M and S appear to have fixed derived nucleotide differences. Second, the fixed differences that characterize the forms are not a result of local adaptation but seem to be integral markers of M and S as incipient species, because they were common to samples from different parts of West Africa and across both sides of the Rift Valley in East Africa. In other words, these differences seem to mark a block of coadapted genes associated with the speciation of M and S. Support for this proposal comes from analysis of the reference *A. gambiae* genome, assembled from the PEST strain. This strain had its origin in the crossing of wild *A. gambiae* from western Kenya (S form) with an *A. gambiae* colony from Nigeria (M form), followed by repeated outcrossing to mosquitoes from Kenya (S form) such that ≈75% of the resulting gene pool was estimated to have been derived from S (42). Nevertheless, the PEST strain is defined as M based on its ribosomal DNA genotype, and this study has shown that, at every locus where there was differentiation between the forms, the PEST sequence clustered with wild M form sequences. This finding suggests that, despite likely homogenization between M and S elsewhere in the genome, the pericentromeric region of the X chromosome has retained the coadapted block of genes that defines the M form. Finally, we note that M and S are sympatric in southern Cameroon, yet in this region of West Africa both forms remain reproductively isolated and both are monomorphic for the standard (uninverted) arrangements of the inversion systems on chromosome 2 (9). Thus, it is conceivable that reduced recombination at the proximal end of the X chromosome is responsible for maintaining ecological and reproductive differences in lieu of alternative chromosomal arrangements.

The genetic architecture of divergence in divisions 5 and 6 deserves comment. It seems unlikely that there are any parts of division 6 that have not been affected by selection. However, loci surveyed in division 5 do not present such consistently high levels of divergence and low amounts of polymorphism (Fig. 2). Microsatellite markers showed divergence between forms only in subdivisions 5C-6 (15), whereas sequence divergence extended distally to 5A. It is possible that microsatellites in 5A-B at one time reflected this wider footprint but subsequently recovered diversity and converged onto overlapping allelic distributions. More likely are two alternative explanations that are not mutually exclusive. Some undifferentiated loci may still have been under selection because they are characterized by very low levels of polymorphism [loci 07907(5A), 15292(5C), 04154(5D), and 13589(5D)]. These loci may have lost polymorphism as a result of the same selective sweeps that affected other loci but by chance were not fixed for alternative neutral variants. However, 01618(5B) and 15233(5B) show a combination of relatively high polymorphism and low divergence, perhaps indicating that these loci were not affected by selection. This interpretation implies multiple targets of selection (at least two) in division 5 and 6. At present, the data do not allow us to suggest the number or position of these targets. Genome scans such as this one, in which patterns of reduced polymorphism and heightened differentiation guide identification of regions associated with adaptive changes, have been successful in a number of organisms, even if the adaptive traits are unknown (ref. 43 and references therein). This strategy is particularly effective when aided by low rates of recombination. However, once a candidate region has been identified, this approach is much less efficient at narrowing down the number or position of candidate genes, especially in the presence of reduced recombination. Further dissection of this region will require alternative approaches, including studies of gene expression and the pattern of replacement and silent substitutions in coding regions. Ultimately, functional studies will need to be devised to link specific genetic traits to the alternative ecological and behavioral characteristics of the molecular forms of *A. gambiae*.

These results are consistent with a divergence-with-gene-flow model (17) for incipient speciation in *A. gambiae*: There are islands of differentiation ("speciation islands") maintained by selection, among them the proximal end of the X chromosome, despite much of the rest of the genome remaining undifferentiated because of ongoing gene flow, as evidenced by introgression of an insecticide resistance gene from S into M (44). These results are also congruent with the findings from a recent study that used the *A. gambiae* Affymetrix GeneChip microarray to examine genome-wide nucleotide divergence between M and S in predicted coding regions; among only three regions identified as significantly diverged (on chromosomes 2R, 2L, and X), the latter two were adjacent to centromeres (14). The GeneChip approach provides a swift and powerful whole-genome overview, albeit with some sacrifice of resolution. Our findings based on sequence analysis suggest that the 566-kb X chromosome island identified by Turner *et al.* (14) may be underestimated by an order of magnitude.

Theory predicts that the speciation islands differentiating M and S are likely to harbor genes responsible for divergent ecological adaptations and premating reproductive isolation. Although debate about the taxonomic status of the two molecular forms is unlikely to be resolved any time soon (12), our evidence suggests that they have been evolving under divergent selection pressures. In practical terms, the failure to treat M and S as separate entities could result in misleading models of malaria transmission and erroneous descriptions of mosquito population dynamics. Moreover, the genetic flexibility inherent in *A. gambiae* that allows it to make more efficient use of heterogeneous and changing environments is not only a fascinating scientific problem but a serious public health problem as well. The exploitation of rice fields by the M form increased malaria transmission spatially and temporally, and the expected reduction in competition with S likely increased the vectorial capacity of both forms through increased longevity and density (33). It is hoped that the recently approved initiative to sequence the genomes of M and S separately (www.genome.gov/15014493) will provide much-needed tools to enhance our understanding of speciation in these mosquitoes as well as tools to reduce or eliminate their ability to transmit disease.

We thank M. Kidwell, M. Hahn, and three anonymous reviewers for useful comments. This work was funded by National Institutes of Health Grants AI44003 and AI63508 (to N.J.B.). A.D.S. was supported by a University of Notre Dame Graduate Fellowship from the Arthur J. Schmitt Foundation.

1. Ortiz-Barrientos, D., Reiland, J., Hey, J. & Noor, M. A. (2002) *Genetica* **116,** 167–178.
2. Butlin, R. K. (2005) *Mol. Ecol.* **14,** 2621–2635.
3. Rieseberg, L. H. (2001) *Trends Ecol. Evol.* **16,** 351–358.
4. Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 12084–12088.
5. Navarro, A. & Barton, N. H. (2003) *Science* **300,** 321–324.
6. Ayala, F. J. & Coluzzi, M. (2005) *Proc. Natl. Acad. Sci. USA* **102,** Suppl. 1, 6535–6542.
7. Besansky, N. J., Krzywinski, J., Lehmann, T., Simard, F., Kern, M., Mukabayire, O., Fontenille, D., Toure, Y. T. & Sagnon, N. F. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10818–10823.
8. Slotman, M., Della Torre, A. & Powell, J. R. (2004) *Genetics* **167,** 275–287.
9. Della Torre, A., Tu, Z. & Petrarca, V. (2005) *Insect Biochem. Mol. Biol.* **35,** 755–769.
10. Tripet, F., Toure, Y. T., Taylor, C. E., Norris, D. E., Dolo, G. & Lanzaro, G. C. (2001) *Mol. Ecol.* **10,** 1725–1732.
11. della Torre, A., Costantini, C., Besansky, N. J., Caccone, A., Petrarca, V., Powell, J. R. & Coluzzi, M. (2002) *Science* **298,** 115–117.
12. Gentile, G., Della Torre, A., Maegga, B., Powell, J. R. & Caccone, A. (2002) *Genetics* **161,** 1561–1578.
13. Gentile, G., Santolamazza, F., Fanello, C., Petrarca, V., Caccone, A. & della Torre, A. (2004) *Insect. Mol. Biol.* **13,** 371–377.
14. Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. (2005) *PLoS Biol.* **3,** e285–e318.
15. Stump, A. D., Shoener, J. A., Costantini, C., Sagnon, N. & Besansky, N. J. (2005) *Genetics* **169,** 1509–1519.
16. Barnes, M. J., Lobo, N. F., Coulibaly, M. B., Sagnon, N. F., Costantini, C. & Besansky, N. J. (2005) *Insect Mol. Biol.* **14,** 353–363.
17. Machado, C. A., Kliman, R. M., Markert, J. A. & Hey, J. (2002) *Mol. Biol. Evol.* **19,** 472–488.
18. Nachman, M. W. (2002) *Curr. Opin. Genet. Dev.* **12,** 657–663.
19. Gillies, M. T. & De Meillon, B. (1968) *The Anophelinae of Africa South of the Sahara* (S. Afr. Inst. Med. Res., Johannesburg).
20. Fanello, C., Santolamazza, F. & della Torre, A. (2002) *Med. Vet. Entomol.* **16,** 461–464.
21. Rozen, S. & Skaletsky, H. J. (2000) in *Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, eds. Krawetz, S. & Misener, S. (Humana, Totowa, NJ), pp. 365–386.
22. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997) *Nucleic Acids Res.* **24,** 4876–4882.
23. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003) *Bioinformatics* **19,** 2496–2497.
24. Schneider, S., Roessli, D. & Excoffier, L. (2000) ARLEQUIN, A Software for Population Genetics Data Analysis (Univ. of Geneva, Geneva), Version 2.0. .
25. Wright, S. I. & Charlesworth, B. (2004) *Genetics* **168,** 1071–1076.
26. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17,** 1244–1245.
27. Hudson, R. R. & Kaplan, N. L. (1985) *Genetics* **111,** 147–164.
28. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
29. Hudson, R. R., Slatkin, M. & Maddison, W. P. (1992) *Genetics* **132,** 583–589.
30. Hudson, R. R., Kreitman, M. & Aguade, M. (1987) *Genetics* **116,** 153–159.
31. Rozas, J., Gullaud, M., Blandin, G. & Aguade, M. (2001) *Genetics* **158,** 1147–1155.
32. Tamura, K., Subramanian, S. & Kumar, S. (2004) *Mol. Biol. Evol.* **21,** 36–44.
33. Coluzzi, M., Sabatini, A., Della Torre, A., Di Deco, M. A. & Petrarca, V. (2002) *Science* **298,** 1415–1418.
34. Smith, J. M. & Haigh, J. (1974) *Genet. Res.* **23,** 23–35.
35. Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993) *Genetics* **134,** 1289–1303.
36. Stajich, J. E. & Hahn, M. W. (2005) *Mol. Biol. Evol.* **22,** 63–73.
37. Wiehe, T. (1998) *Theor. Popul. Biol.* **53,** 272–283.
38. Payseur, B. A. & Nachman, M. W. (2000) *Genetics* **156,** 1285–1298.
39. Tenaillon, M. I., Sawkins, M. C., Anderson, L. K., Stack, S. M., Doebley, J. & Gaut, B. S. (2002) *Genetics* **162,** 1401–1413.
40. Fernando Vazquez, J., Perez, T., Albornoz, J. & Dominguez, A. (2000) *Genet. Res.* **76,** 323–326.
41. Nurminsky, D. I. (2001) *Cell Mol. Life Sci.* **58,** 125–134.
42. Githeko, A. K., Brandling-Bennett, A. D., Beier, M., Atieli, F., Owaga, M. & Collins, F. H. (1992) *Trans R. Soc. Trop. Med. Hyg.* **86,** 355–358.
43. Storz, J. F., Payseur, B. A. & Nachman, M. W. (2004) *Mol. Biol. Evol.* **21,** 1800–1811.
44. Weill, M., Chandre, F., Brengues, C., Manguin, S., Akogbeto, M., Pasteur, N., Guillet, P. & Raymond, M. (2000) *Insect Mol. Biol.* **9,** 451–455.